情報量の巨大化と情報の質へのアプローチ：現状と課題

情報量の巨大化と情報の質へのアプローチ：現状と課題

二瓶　真理子

1　デジタルユニバースの急拡大と情報のビッグデータ化

二〇一五年のある調査によれば、全世界のツイッターユーザによる一分間当たりのツイート数は347222回、またユーチューブユーザがアップロードする動画は一分間当たり三〇〇時間、そしてフェイスブックユーザの「いいね！」は一分間当たり4166667回にのぼるという[1]。これらアクションによって世界中で膨大な量のデータが生み出され、その総量は刻一刻と拡大し続けている。

データ生成量の拡大の勢いは年々高まっており、二〇二〇年までに全世界で生成されるデジタルデータの総量は40ゼタバイトを超えると予測されている[2]。この予測は、米調査会社IDC（Internet Data Corporation）が二〇一二年十二月に発表したものである。IDCは二〇〇七年以来、全世界で生成または複製されたデジタルデータの総量を「デジタルユニバース（DU）」と名付け、その現状と推移予測を定期的に発表してきている。二〇〇五年には0・13ゼタバイトであったDUは、二〇一〇年には1・3ゼタバイト、二〇一二年には2・8ゼタバイトまでに拡大している[3]。IDCは、二〇〇五年から二〇二〇年までにDUは三〇〇倍に増大、また二〇一二年から二〇二

○年までは二年ごとに二倍増のペースで拡大が続くと予測している。この予測に基づくと二〇二〇年のＤＵは４０ゼタバイトを超えることになる。[4] また、米通信会社シスコによれば、全世界のＩＰトラフィック量についても二〇一六年末には年間１・１ゼタバイトに達すると見込まれている。[5] 我々は人類史上初めて、ゼタバイト時代を生きる世代なのである。

ビッグデータ（big data）という語はもともとコンピュータ業界で一九九〇年中頃から大きなデータセットの処理や分析という意味で使用されていたようだが、ＤＵがゼタバイトの大台を超えた二〇一〇年代に入って急激にバズワード化した印象がある。おそらく最も流通している３Ｖ定義によれば、ビッグデータとは、巨大な量のデータから成る（Volume）、リアルタイムで生成される（Velocity）、構造化データだけでなく非構造化・半構造化データも含む（Variety）という三つの特性を持つとされる。[6]

冒頭でもふれたユーチューブやブログ等は、写真や動画、絵文字などの非構造化データをリアルタイムに生成しており、ビッグデータの代表的な生成源のひとつである。だが、ブログのように能動的に生成されるデータだけでなく、デジタルシャドウとも言われる人の移動や行動に付随して自動的に生成されるデータ（POSデータ、RFIDデータなど）も大きな割合をしめる。[7] ＩＤＣは、二〇〇七年の時点でＤＵの半分以上はデジタルシャドウだとしている。また、街頭や工場などに設置された監視カメラや気象用センサーなどから送られてくる観測データのような機械生成データも年々増加してきており、二〇二〇年には全生成量の４０パーセントが人間の行動とは関係を持たない機械生成データになるとの予測もある。[8] ビッグデータとは、多様な生成源と多様な構造を持った膨大な量のデータが混在したデジタルユニバースなのである。

最近では、買い物や医療など身近な場面でもビッグデータという言葉を目にする。しかし、ビッグデータの利活

用は現状では模索段階であり[9]、それによって我々が受けるかもしれない恩恵も明確でない。ビッグデータの定義や活用法の議論状況も未成熟であるなかで、本稿ではまず現状の把握と問題の整理を目的として議論をすすめたい。ビッグデータブームの背景には、先に述べたデジタルデータ量の急拡大がある。以下ではまず、ＩＴＣの進歩と浸透に伴って発生してきたこの現象が、我々にどのような事態をもたらそうとしているのかをみる。その後、情報の量ではなく質への視点として、「情報の質（information quality）」概念を定義し評価する研究プログラムの動向を紹介する。これにより、データ量の巨大化という問題の現状と、それに対して我々が取り組むべき課題を明らかにしたい。

2　情報量巨大化という問題

（1）情報オーバーロード問題との関係

さて、我々がゼタバイト時代を生きていることはすでに見たが、その我々にとって情報量が巨大化することの何が問題なのか。多すぎて困ることが問題なのだろうか。

確かに、二〇〇〇年代にクローズアップされた情報オーバーロード（information overload）は、現代の大きな社会問題のひとつである。情報オーバーロードとは、たとえば「ひとりの人物が引き受けられる量を超えた情報」とか、「意思決定をするために必要な分（または自分の持ち時間中に理解し消化することができる分）よりも多くの情報を受信することと、それに対して時代遅れの管理方法で対処しようとすることによるストレス」とされる（J. B. Strother etal (eds.) [2012] pp.1-2）。要するに、自身の処理許容範囲を超えた情報量にさらされる状態や、そのことによる心理的、組織的、経済的弊害が生じているとき、我々は情報オーバーロード状態にある。とくにビジネス

面で、過多のメール送受信やSNS利用によって従業員の生産性や意思決定能力が下がることが問題視されており、情報オーバーロードが二〇一一年にアメリカ経済に与えた損出額は約九百億ドルであるとの試算もある[10]。

情報オーバーロードと情報量の巨大化は、ITCの普及という同じ背景を持つこともあり、同種の問題とみなされがちである。しかし、これらは関連し合うとしても独立の問題である。現在の情報倫理・情報哲学を牽引するルチアーノ・フロリディは、情報オーバーロードは食べ切れるよりも多くの量を無理に喉奥に流し込まれている状態であるのに対して、ビッグデータ問題はいままで食べることができたよりも多くのメニューを提供するパーティ会場でぼやいているような違いがあると述べている（Floridi [2014] p.15）。ビッグデータの潜在的有効性についてかなり楽観的である点は気になるが、要は、情報の巨大化それじたいが我々に情報の消費を強いるわけではないということだろう。

この点は、前節でみた調査結果とも符合する。情報オーバーロードは主に人間が能動的に生成した電子メールやSNSデータによって生じる問題である。だが、デジタルユニバースの大部分は人間が直接的に意識しないデータシャドウや機械生成データで占められているのだった。これらのデータは、我々にその処理を求めることもなく、我々の与り知らぬ所で増えていくだけである。ゼタバイト規模の情報量巨大化は、自分の処理許容量以上の情報を食わされて辟易するという意味での〝多すぎて困る〟情報オーバーロード問題とは別のものなのである。

そうだとすると、しかし、情報量の巨大化は実は情報オーバーロードよりもごく単純な、技術的に解決できる問題であるようにも思われる。データシャドウの生成や処理については人間の認知能力の限界を考慮にいれなくてよいのだとしたら、その種のデータ量がいかに巨大化するとはいえ、技術的な面でのデータ処理能力を上げていけばよい話なのではないか。だが、ことはそれほど単純ではない。

（2）デジタルストレージの量的限界と質的限界

IDCの二〇〇八年報告によると、DUは二〇〇七年に281エクサバイトに達し、以来全世界のデータストレージの総容量を上回って拡大している。データ生成量の拡大に比べるとあまり注目されていないが、データ生成量とストレージ容量とのギャップは、一年ごとにおよそ二倍のペースで広がるとされ、二〇一一年時点で生成データの約半数が保存先を持たないと試算されている[11]。現在生成されているデータの大半が、恒久的な保存先を持たずいずれ捨てられる運命にある。

フロリディは我々に向けて、上のような量的限界とともにデジタルストレージの質的限界への注目も促している（Floridi [2014] p.18）。我々の記憶と記録の媒体は、口頭伝承から文字文書へ、そしてデジタルデータへと移り変わってきた。量の増大とともに、安定性と永続性も高まったはずだと我々は考えている。だが、そのイメージとは逆に、デジタル記録は口頭伝承と同じくらいに不安定で忘却性が高い。フロッピーディスクやミニディスクに保存された記録を今すぐ取り出せるだろうか。WWWのリンク切れエラーがどれだけ多いことか[12]。

デジタル記録の不安定性は、そう遠くない将来世代にとって歴史の消失をもたらしかねない。WWWが典型的であるが、ドキュメントの保存は古いバージョンへの上書きであり、原理的には何度でも可能なリライトを繰り返すうちに過去の記録の歴史的堆積は我々からは見えなくなっていく。また、現代の主要記録媒体の変遷スピードは、今後ますます多くの旧式媒体内部のデータがアクセス不能なものとなっていくことを予想させる。我々の知識の大部分がこのような忘却性の高い不安定な場に置かれていることは、我々が無時間的な永遠の現在の監獄に自ら引きこもっていくことでもあるとフロリディは警鐘を鳴らす。

ストレージ容量はすでに限界を迎えているため、限られた容量内部で情報の質的水準を維持するためには、現行

の情報処理や記録蓄積方式に代わる対策や制度が必要となる。情報量の拡大現象は、既存技術の量的拡大だけでは対応できない。デジタルシャドウやセンサデータなど我々に直接見えてこないデータも含めたデジタルユニバース全体に対して、どのデータをどのような形で残すのか、どのデータを捨てて保存容量を確保するのか。デジタルストレージの量的・質的限界が顕在化した今求められているのは、情報の内容やニーズに基づくキュレーションである。

歴史保存の問題に対しては、持続可能なデジタルアーカイブのための取り組みが世界各地で進められている。例えば、二〇〇三年に設立された国際インターネット保存コンソーシアム（IIPC）は、保存形式や保存規準の規格標準化を目指したウェブアーカイブの国際連携機関であり、日本の国会図書館も含め四十五以上の国立図書館が加盟している[13]。

しかし、データや情報の内容の評価は、量的な処理よりもはるかに複雑な問題である。コンピュータ科学、情報理論、図書館情報学、哲学など様々な分野で様々な側面から情報の「質」へのアプローチが模索されているが、統一的な見解は存在していない。以下では、それらアプローチのひとつである「情報質（information quality）」をめぐる研究プログラムの概要と近年の情報巨大化に対応した動向をみてみたい。このプログラムを取り上げるのは、一定の理論体系と実践実績をもつ研究プログラムとして一定期間認知されていることと、このプログラムの変遷と動向がデータ量急拡大現象の内実とそれに対する対策を整理するためのよい手掛かりになると考えられるからである。

3　情報の質へのアプローチ

（1）ＴＤＱＭプログラム：目的への適合としてのデータ品質

デジタルデータの質ないし品質（quality）を指す「データクオリティ（data quality/DQ）」または「インフォメーションクオリティ（information quality/IQ）」という語は、コンピュータ科学、統計学、経営学等にまたがる学際的研究の一分野として欧米で認知されてきている。[14] ここ二十年ほどで三つの学術雑誌、the Data Quality Journal (1995-)、theInternational Journal of Information Quality (2007-)、the ACM Journal of Data and Information Quality (2008-) が創刊されたほか、実務家や研究者が一堂に集う国際会議 International Conference on Information Quality も一九九六年以来毎年開催されている。

一連のデータクオリティ研究の先駆けとなったのが、ＭＩＴスローン経営学大学院のリチャード・ワンを中心に一九九〇年代から展開されたＭＩＴトータルデータクオリティマネジメントプログラム（ＴＤＱＭ）である。[15] 職場のＩＣＴ化も情報オーバーロードも顕著な問題ではなかった当時、ＴＤＱＭが目指したのはデータの品質管理であった。一九八〇年代のアメリカでは製造業を中心に総合品質管理（ＴＱＭ）フレームワークの導入が進められており、ＴＤＱＭはとくにデータベース製品に特化した総合品質管理フレームワークの確立を目標としていた。ここではデータクオリティという概念は、顧客満足度の高いデータ製品生産に向けた組織管理と改善という文脈のなかで文字通り「データ品質」として扱われていた。

当然、現在のＤＱ／ＩＱ研究を支えるモチベーションは品質管理に限定されない。二〇〇〇年代に入って学会誌創設が相次いだことは、情報オーバーロードやビッグデータブームとの関連でＤＱ／ＩＱ概念が改めて注目されたことによるのだろう。だが、背景にある問題は変化してきたものの、ＴＤＱＭの方法論や定義は現在でも初期の雛

カテゴリー	次元
固有 IQ	正確性、信頼性、客観性、評判
文脈 IQ	付加価値、適時性、関連性、適量性、完全性
表象 IQ	解釈可能性、理解の容易性、表現の一貫性、表現の簡潔性
アクセス IQ	セキュリティ、アクセスの容易性

図1　DQ/IQ評価のためのカテゴリーと次元（Wang & Strong「1996」）

形として言及されることが多い。

DQ／IQ概念の初期の体系的明確化とされる論文において、ワンはデータおよび情報の質は「目的または用途への適合」によって決まると定義し、当該データの目的への適合度を評価するための十五の次元（dimensions）と各次元に対応する定量的計測方法を提示した（Wang & Strong [1996]）。正確性、完全性、信頼性、適時性などの次元は、データを扱う実務家からのインタビューから経験的に抽出されたもので、データそのものが持つ性質にかかわる固有IQ、当該のタスクや目的への適合にかかわる文脈IQ、データの可読性や表現形式にかかわる表現IQ、データのアクセス方法とセキュリティにかかわるアクセスIQという四つのカテゴリーに分類される（図1参照）。

ここでワンが強調したのは、データの「よさ」は多元的であり、その正確性のみには還元できないという点である。いくら正確でも古い情報は役立たないかもしれない。また、広く浅くデータを集めたい場合もあれば、狭く深く知りたい場合もある。データがよい品質であるか否かの評価は、その用途や目的、評価者によって変わりうるため規準の定式化が難しい。ワンは、データの次元を多元化し、用途や目的あるいは評価者ごとにそれぞれの次元への重み付けが変化することを許容することで、目的相対的かつ客観的で統一的な評価枠組みを提示したといえる。目的への適合を多次元から評価するというワンのフレームワークは、顧客のニーズを、データベース構築者、管理保守者が理解するという当初のTDQM内部での目的に大きく貢献したほか、データベース評価の方法論として一定の影響力を持った[16]。

これ以降データベース以外も含めた情報材のＤＱ評価を目指す類似研究が数多く発表された。しかし、これらが統一的な方法論理論とか技術的ツールに収束していくことはなかったようだ。エプラーは、一九九〇年代以降に発表された数十のＤＱ／ＩＱフレームワークモデルを比較し、この分野の全般的問題として、評価次元のセット及び各次元の計測手法が収束しないこと[17]、どのモデルも対象とするデータ種（テキスト、ウェブページ、データベースなど）に特化しており汎用性が低いこと、ソフトウェアなど具体的な技術への展開が進まないことを挙げている（Eppler [2006]）。当然のことではあるが、対象とする領域やデータ形式に対しての依存性が強いモデルほどその対象領域の評価には優れるが、汎用性は低くなる。一九九〇年代後半から急速に進んだＩＣＴ化は、ＤＱ／ＩＱ研究に情報オーバーロードの解決という新たな大命題を与えたが、様々な機器やデータ形式の導入により、ＤＱ／ＩＱモデルを細分化させ統一的方法論の構築をますます難しくしたとも言える。

（2）ビッグデータの質評価

さらに二〇〇〇年代からのデジタルユニバースの急拡大は、ＤＱ／ＩＱ研究への期待と注目を増大させたと同時にＤＱ／ＩＱ研究のコアな前提のひとつを掘り崩しもした。ＴＤＱＭが主対象としていたのは、静的かつ構造化されたデータ、典型的には関係データベースであったが、現在のＤＵ内部でその種のデータが占めるのはごく一部に過ぎない。先に触れたようにＤＵの大部分は、人間の行為に付随して生じるデジタルシャドウとセンサー等から自動的に収集される機械生成データである。従来のＤＱ／ＩＱ概念にとって最も大きな問題となるのは、これらデータがその生成に先立って特定の用途や目的を持たない点である。使用用途への適合という観点からの質評価をそのままこれらに当てはめることはできない。

また、人間が生成するデータについても新しい課題が生じている。ソーシャルデータとも言われるブログ、ＳＮＳ等のデータは、画像、映像、テキスト、メールなど様々なデータ種を含む。これらデータの大部分は非構造化データないし半構造化データであり、データベースのような形式が標準化された構造化データと同様の質評価次元や手法は適用できない。

二〇〇〇年前後にＴＤＱＭプログラムの方法論的精緻を目指しイタリアでComplete DQMプログラムを牽引していた「イタリア学派」のバティーニらは、目的の不在とデータ種の多様化に直面して、ＤＱ／ＩＱ研究は現在パラダイムチェンジを迎えていると述べる（Batini et al [2014]）。この転換をイタリア学派は、データクオリティからインフォメーションクオリティへの移行として総括している。初期のプログラムは、構造化されたデジタルデータ製品の品質評価という特殊事例を典型的課題としていたが、今後は多様なデータ種がもつ情報の質の概念化というより一般的な視点からＤＱ／ＩＱ研究を進めていかなくてはならない[18]。

イタリア学派とその周辺の最近の動向から、ビッグデータＩＱ研究の現状と課題を見ておこう[19]。ビッグデータには様々な生成源を持つ多種のデータが混在しており、そのＩＱはそれぞれの生成源や種類に依存して考慮される必要がある。UNECE（United Nations Economic Commission for Europe）は、ビッグデータを、人間生成データ、プロセス媒介データ、機械生成データの三つに区分している（図2参照）。近年の研究プログラムは、この生成源区分とデータ種区分に従って、従来的なＤＱ／ＩＱ概念を拡張しつつ、各データタイプに適合したＩＱの概念化と定量的計測法の開発を目指している。

プロセス媒介データは、個々のデータがウェブ上などに分散、重複しており、実世界のプロセス解析に使用するにはデータ統合が必要となる問題がある。だが、医療や購買など予め特定の対象領域が決まっており、それぞれ規

生成源	主なデータ	データ構造	人間の介在	*他の名称
人間生成	テキスト、画像、映像、SNSなど	非構造化、半構造化データ	直接的	ソーシャルデータ
プロセス媒介	POS、RFID、信用情報、医療情報など	構造化データ	間接的	デジタルシャドウ
機械生成	センサデータ、ログ、監視データ、GPSなど	構造化データ	なし	実世界事実データ

図2　生成源によるビッグデータの区分（UNECE Classification of types of data 2015）
*他の名称は、筆者が付加

格標準化された構造を持ちメタデータ付与もある程度標準化されているため、従来のデータベース研究からの応用が最も容易と見込まれているようだ。じっさいに、実践面での解析や利活用が最も進んでいるのもこのタイプで、購買レコメンドシステムや配送物の追跡システムなどに使用されている。従来の正確性、一貫性、完全性のほか、データの冗長性、有効期限性などが重要な評価次元になってくると議論されている。

今後ますますの増加が予想される機械生成データは、データ構造の点では従来の技術で問題なく処理できる。また、実世界の出来事や状況を記録・測定した事実についてのデータであるので、そのIQ評価次元もある程度収束しやすいようだ。正確性、一貫性、完全性に加え、生成源の物理的な信頼性とデータの新鮮度が主要次元として提案されることが多い。この種のデータは、科学的測定データとしての他、ある場所がある時刻に何度であったとか、ある人物がある時刻にある場所にいたといった「事実」を示すデータとして、工場の異常感知や犯罪者の追跡などへの活用が目指されている。だが、測定機器の設置の仕方や、環境条件・気象条件による故障やエラーの多さも機械生成データの特徴であるため、事実の証拠としてデータを使用するためには、機器状態や機器周辺の物理的環境への信頼性評価が必須の項目となる。機械生成データの最大の問題は、その量とスピードである。原理的には可能でも、膨大な量をリアルタイムで自動処理、評価する

ことは現時点の技術だけでは困難であり、新しい技術基盤の開発が待たれる。また、すべての生成データの保存は現実的に不可能であるから、データ廃棄の規準とシステム化も必要であろう。

最後に、人間生成データだが、この種のデータについてのＩＱ研究が最も難しく蓄積も少ない。提案されるＩＱ評価次元についても現時点では収束は見られない[20]。テキスト、画像、映像等の多様なデータ種ごとにＩＱが議論されているが、これらデータ種はどれも標準化された形式がなく、メタデータも付与されていないことが多い。ソーシャルデータの利活用は大きな期待を持たれているが、現時点ではその大部分が機械による自動的処理が不可能な形式のまま存在している状態である。

ソーシャルデータの「質」を考えるさいにもうひとつ問題となるのは、データの対象つまり情報の内容が必ずしも正確性を志向しない点だろう。他の生成源データは、それにすべて還元されるわけではなくとも正確性という評価次元を必要条件とするが、ソーシャルデータはその限りではないかもしれない。機械生成データの対象は実世界の事実であるが、ＳＮＳやブログ、インターネット上の内容は誤りも含まれるし、必ずしも真偽を持たない内容であることもある。前半でみたウェブページのアーカイブ化の問題とも関連するが、情報生成源に対する信頼性（筆者、作成日時・場所、情報媒体の形式、改変歴など）と、情報内容の妥当性という意味での信頼性（正誤、エビデンスの有無、評判など）を評価する手法が必要となるだろう[21]。

以上データ生成源ごとに状況をみてきたが、基本的には、それぞれのデータの対象領域や特徴から潜在的用途を割り出し、その用途にとって必要と考えられるＩＱ評価次元を設定する方針が採られている。データベースの品質評価フレームワークに比べるとまだまだプロトタイプ以前の状態ではあるが、プロセス媒介データと機械生成データについては質評価の見通しは徐々につきつつあるようだ。しかし、量とスピードに対する技術的問題と、人間生

成データの評価についての概念的問題は現時点ではほぼ手付かずのままである。

4　デジタルユニバースの質的キュレーションに向けて

ここまで、前半では、生成情報量がストレージ容量を超える勢いで急拡大を続けていること、限られた容量内部で我々の知識を保存していくためには情報の取捨選択が必須であることをみた。後半では、情報の質を評価するＤＱ／ＩＱ研究プログラムの近年までの動向を追ってきた。本稿で取り上げたＤＱ／ＩＱ概念が情報の質についての唯一の定義ではなく、また、情報質評価が直接的に取捨選択規準を与えてくれるわけではない。だが、情報が持つ質の多元性に注目した情報評価という枠組みは、今後、取捨選択の基準を議論するさいの有効な視点となりうる。また、膨大な情報量のすべてを人間の手でキュレーションすることはほぼ不可能であるから、各質評価次元に対して定量的な計測法を対応させることで、情報の質の機械的自動評価を可能にしている点も参考になる。とくに、ＤＵの大部分を占める機械生成、プロセス媒介データについては、定量的な質評価に基づいたデータ廃棄、情報キュレーションの規準や方式を構築していくことも可能かもしれない[22]。

しかしながら、先ほどみたように、非構造化データの多様化と増大に対する対策は立ち遅れている。人間によって生成されるソーシャルデータについての分析は、ＤＱ／ＩＱプログラムを含め、ほぼ進んでいない状態である。ＤＵ全体からみればごく一部に過ぎないが、ソーシャルデータは、我々の眼にみえている情報領域であり、情報オーバーロード問題として我々に直接弊害をもたらす側面ももつ。また歴史的記録の消失の問題も無視できない。データ構造の標準化の問題と情報内容の評価と保存の問題の両方について、工学的・技術的知見と人文科学、社会科学の知見を持ち寄り、取り組んでいく必要があるだろう。

情報量の巨大化それじたいに対して、現時点で楽観的にも悲観的にもなりすぎる必要はない。だが、常に手持ちの技術や知識の現状から飛躍せずに、巨大なデジタルユニバースの中に我々が求めている価値は何なのか、そこに何を残したいのかを冷静に見極めていくことが、ゼタバイト時代を生きる我々に求められている。

註

＊以下で指示したURLはすべて二〇一六年八月九日時点のものである。

（1）米に本社をおく調査会社DEMOによる二〇一五年八月の報告「Data Never Sleeps 3.0」による。https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/ なお、DEMOは、二〇一六年六月に最新版「Data Never Sleeps 4.0」を公開している。https://www.domo.com/blog/2016/06/data-never-sleeps-4-0/

（2）THE DIGITAL UNIVERSE IN 2020：Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East（IDC 2012.12）https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

（3）ゼタは10^21を表し、データ量の単位としては、キロ、メガ、ギガ、テラ、ペタ、エクサの次に当たる単位である。1ゼタバイトは10^12ギガバイト。Ciscoのインフォグラフィック「the internet 2015 is the Dawn of the Zettabyte Era」によると「1ギガバイトがコーヒー1杯（約330cc）だとすると、1ゼタバイトは万里の長城の体積に相当する」。http://blogs.cisco.com/news/the-dawn-of-the-zettabyte-era-infographic

（4）二年で倍増ペースで単純計算すれば二〇二〇年には44ゼタバイトということになる。同じ拡大ペースが続いていけば、二〇三〇年までにDUは約1・3ヨッタバイトに達する。ヨッタは10^24を表すゼタの次のデータ量単位である。

（5）二〇一六年六月発表「Cisco Visual Networking Index（VNI）：予測と方法論2015-2020」による。http://www.cisco.com/web/JP/solution/isp/ipngn/literature/white_paper_c11-481360.html

（6）ビッグデータの3V定義が初めて明文化されたのは以下といわれる。Laney, D [2001] 3D data management: Controlling data volume, velocity and variety. Mata Group.

（7）日本国内についての別の調査にはなるが、平成二十五年総務省情報通信白書（第一部第三節「ビッグデータの活用が促す成長の可能性」）によると、日本国内のビッグデータ全流通量のうち最も大きな割合を占めるのがPOSデータ（物品売上げデータ）、次いでICチップが埋め込まれた乗車カード、社員証等から生成されるRFIDデータ、GPSデータ（現在位置データ）と続く。ブログ・.

SNS と比較して、POS データはその約六七五倍、RFID データは約五一五倍の量のデータが流通している計算になる。http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc113220.html

（8）前掲 IDC2012報告書「THE DIGITAL UNIVERSE IN 2020」による。

（9）前掲 IDC2012報告書では、二〇一二年の全DUのうち分析すれば有効な情報が取り出せるのは23%であるのに対し、分析に必要な付加情報のタグ付けをされたデータは約3%、実際に分析されたデータは約0・5%に過ぎないと報告している。

（10）J.Spira, "Information Overload: How $900 Billion - What is Your Organization's Exposure?" http://www.basexblog.com/2008/12/19/information-overload-now-900-billion-what-is-your-organizations-exposure/

（11）IDC2008報告書「The Diverse and Exploding Digital Universe」による。http://www.atour.com/media/images/service/IDC-EMC-The-Diverse-and-Exploding-Digital-Universe-2008.pdf

（12）URL の平均寿命は四四―七五日との報告もある。Marieke Guy, "What's the average lifespan of a Web page?…or is it easier to ask how long is a piece of string?" 2009.8.12° http://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page/

（13）国立国会図書館のサイト「カレントアウェアネス・ポータル」で世界各地のデジタルアーカイブの最新動向を知ることができる。http://current.ndl.go.jp/

（14）quality という語は品質とも質とも訳せる。以下では、研究の内容によって同じ語を「品質」と「質」に訳し分けている。また、data (quality) と information (quality) について、MIT プログラムではほぼ同義の互換性のある語としている。だが、後述するイタリア学派では、data は構造化デジタルデータのみを指し、information は非構造化デジタルデータおよび地図や文章が持つ情報内容一般を指すものとして区別している。

（15）MITのTDQMプログラムは後継の MIT Information Quality (MITIQ) Program（http://mitiq.mit.edu/）に吸収される形で二〇〇三年に終了した。

（16）二〇〇八年にデータベース製品のための国際標準規格（ISO/IEC25012:2008 Data Quality）が制定された。ワンらによる多次元質評価の方法と同様にデータ品質評価のための一五次元が規定されている。但し、ISO規格では次元のカテゴリー化はされていない。

（17）次元の収束のなさは、データ種、データタイプの相違のほかでは、次元抽出方法の相違、次元間の優先関係、トレードオフ関係の考え方の相違などによる。次元抽出には、アンケートやインタビューなどの経験的な手法のほか、特定の存在論や認識論に基づく理論的抽出、両者の折衷手法などいくつかの方法が採られている。

（18）ここでまとめた動向は、以下を中心に参照した。Batini and Scannapieco [2016], Firmani et al [2016], Batini et al [2014], Illari & Flirdi [2014]

（19）フロリディも、品質管理という背景にとらわれず、哲学や認識論等の理論的背景からIQを捉える方向を採る。Floridi [2014] は、

自身の情報哲学における抽象レベル理論に基づきIQ概念を再考しようとしている。

（20）諸データ種のうち、画像については画像質評価研究の系譜から比較的研究蓄積が厚い。現在はFUNモデルといわれる忠実性（Fidelity）、有用性（Usefulness）、自然さ（Naturalness）の三次元からの主観的画像評価方法が主流になりつつある。

（21）クラウドソースの集合知情報の信頼性の問題は、ピア共同体内部での科学的データの信頼性構成の問題と同型であり、この種の研究蓄積がIQ研究にも貢献しうることをフロリディらは指摘している。（Illari & Floridi [2014] p.13）

（22）本稿はデータの質評価を主題としたため、データ保存の政治的・倫理的問題には触れずにきたが、誰が情報の取捨選択、廃棄の主導権を持つのか、誰がデータを保管し管理するのかは、DUのキュレーションにとって避けて通れない大きな問題である。

文献

Batini and Acannapieco [2016] *Data and Information Quality: Dimensions, Principles and Techniques,* Springer.

Batini et al. [2014] "Opening the Closed World: A Survey of Information Quality Research in the Wild" in Floridi and Illari (eds.) [2014] pp. 43-74.

Eppler [2006] *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes,* Springer.

Firmani et al. [2016] "On the Meaningfulness of "Big Data Quality" (Invited Paper)", *Data Sci. Eng.* 1 (1):6-20.

Fliridi and Illari (eds.) [2014] *The Philosophy of Information Quality,* Springer Stbthese Library.

Floridi [2014] "Big Data and Information Quality" in Floridi and Illari (eds.) [2014] pp.303-315.

Floridi [2014] *The 4TH Revolution: How the Infosphere is reshaping Human Reality,* Oxford University Press.

Illari and Floridi [2014] "Information Quality, Data and Philosphy" in Floridi and Illari (eds.) [2014] pp.5-24.

Strother et al. [2012] *Information Overload: An International Challenge for Professional Engineers and Technical Communicators,* IEEE Press.

Wang and Strong [1996] "Beyond accuracy: what data quality means to data consumers", *Journal of Management Information Systems,* 12 (4), pp.5-33.

＊本研究は科研費16K20911および東北大学・学際重点研究プログラム「ヨッタスケールデータの科学技術」の助成を受けたものである。

（にくい　まりこ・東北大学電気通信研究所）